EXHIBIT T

TO DECLARATION OF S. MERRILL WEISS IN SUPPORT OF PLAINTIFF ACACIA MEDIA TECHNOLOGIES CORPORATION'S MEMORANDUM OF POINTS AND AUTHORITIES IN OPPOSITION TO ROUND 3 DEFENDANTS' MOTION FOR SUMMARY JUDGMENT OF INVALIDITY UNDER 35 U.S.C. § 112 OF THE '992, '863, AND '702 PATENTS; AND SATELLITE DEFENDANTS' MOTION FOR SUMMARY JUDGMENT OF INVALIDITY OF THE '992, '863, AND '720 PATENTS

Hovsepian v. Apple, Inc.

Doc. 325 Att. 20

Dockets.Justia.com

Predictive Coding of Speech at Low Bit Rates

BISHNU S. ATAL, FELLOW, IEEE

Abstract—Predictive coding is a promising approach for speech coding. In this paper, we review the recent work on adaptive predictive coding of speech signals, with particular emphasis on achieving high speech quality at low bit rates (less than 10 kbits/s). Efficient prediction of the redundant structure in speech signals is obviously important for proper functioning of a predictive coder. It is equally important to ensure that the distortion in the coded speech signal be *perceptually* small. The subjective loudness of quantization noise depends both on the short-time spectrum of the noise and its relation to the short-time spectrum of the speech signal. The noise in the formant regions is partially masked by the speech signal itself. This masking of quantization noise by speech signal allows one to use low bit rates while maintaining high speech quality. This paper will present generalizations of predictive coding for minimizing subjective distortion in the reconstructed speech signal at the receiver. The quantizer in predictive coders quantizes its input on a sample-by-sample basis. Such sample-by-sample (instantaneous) quantization creates difficulty in realizing an arbitrary noise spectrum, particularly at low bit rates. We will describe a new class of speech coders in this paper which could be considered to be a generalization of the predictive coder. These new coders not only allow one to realize the precise optimum noise spectrum which is crucial to achieving very low bit rates, but also represent the important first step in bridging the gap between waveform coders and vocoders without suffering from their limitations.

I. INTRODUCTION

PREDICTIVE coding is an efficient method of converting signals into digital form [1], [2]. The basic idea behind predictive coding is very simple and is illustrated in Fig. 1. The coding efficiency is achieved by removing the redundant structure from the signal before digitization. The predictor *P* forms the estimate for the current sample of the input signal based on the past reconstructed values of the signal at the receiver. The difference between the current value of the input signal and its predicted value is quantized and sent to the receiver. The receiver constructs the next sample of the signal by adding the received signal to the predicted estimate of the present sample.

The properties of speech signals vary from one sound to another. It is therefore necessary for efficient coding that both the predictor and the quantizer in Fig. 1 be adaptive [3]-[5]. The digital channel in an adaptive predictive coding system carries information both about the quantized prediction residual and the time-varying parameters of the adaptive predictor and the quantizer (often referred to as side information). The transmission of the prediction residual usually requires a significantly larger number of bits per second in comparison to

Manuscript received September 9, 1981; revised December 11, 1981. This paper was presented in part at the International Conferences on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978, Washington, DC, April 1979, Denver, CO, April 1980, and Atlanta, GA, March 1981.

The author is with Bell Laboratories, Murray Hill, NJ 07974.

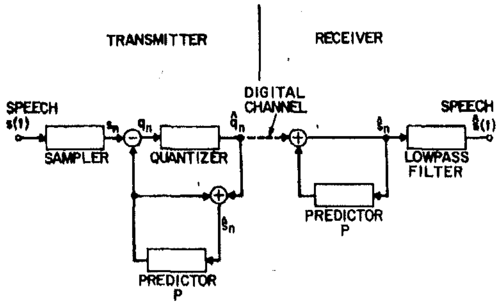


Fig. 1. Block diagram of a predictive coder.

the side information. For example, the bit rate for the prediction residual is 8 kbits/s for speech sampled at 8000 samples/s and the prediction residual quantized at 1 bit/sample. The side information typically requires 3-5 kbits/s.

It can be shown that the quantization noise appearing in the output speech signal is identical to the quantizer error (the difference between the output and the input of the quantizer) [4]. The spectrum of the quantizer error for a multilevel quantizer with finely spaced levels is approximately flat. Thus, the spectrum of the quantization noise appearing in the reproduced speech signal in the coder shown in Fig. 1 is also flat. Recent work on coding of speech signals has demonstrated that "white" quantization noise is not the optimal choice for realizing minimum *perceptual* distortion in the reproduced speech signal [6]-[9]. We discuss in this paper generalizations of the coder shown in Fig. 1 for producing quantization noise of any desired spectral shape.

The assumption that the spectrum of the quantizer error is flat is only true for a multilevel quantizer with small step size. A coarse quantizer with two or three levels is often used for speech coding at low bit rates. The quantizer error for such coarse quantization is not white. Delayed predictive coding (tree coding) methods are then necessary for realizing proper noise spectrum in the reproduced speech signal [10].

Efficient quantization of the prediction residual is essential in achieving the lowest possible bit rate for a given speech quality. At bit rates lower than about 10 kbits/s it is often necessary to quantize the prediction error with only 1 bit/sample (two levels). Such a coarse quantization is the major source of audible distortion in the reconstructed speech signal. Accurate quantization of high-amplitude portions of the prediction residual is necessary for achieving low perceptual distortion in the reproduced speech signal. Improved quantization procedures are therefore necessary for high-quality speech coding at low bit rates. We discuss in this paper methods which allow accurate quantization of the prediction residual when its amplitude is large but also allow encoding of the prediction residual at fractional bit rates (lower than 1 bit/sample).

0090-6778/82/0400-0600$00.75 © 1982 IEEE

Exhibit I Page 400

II. ADAPTIVE PREDICTIVE CODING SYSTEMS WITHOUT NOISE SHAPING

Adaptive predictive coding (APC) systems for speech signals have been discussed extensively in the literature [3]-[5], [9], [10]. We will review their important features briefly.

Selection of Predictor

For speech signals, the predictor P includes two separate predictors: a first predictor P_s, based on the short-time spectral envelope of the speech signal, and a second predictor P_d, based on the short-time spectral fine structure. The short-time spectral envelope of speech is determined by the frequency response of the vocal tract and for voiced speech also by the spectrum of the glottal pulse. The spectral fine structure arising from the quasi-periodic nature of voiced speech is determined mainly by the pitch period and the degree of voiced periodicity. The fine structure of unvoiced speech is random and therefore cannot be used for prediction.

Prediction Based on Spectral Envelope

Prediction based on the spectral envelope involves relatively short delays. The number of predictor coefficients is typically 16 for speech sampled at 8 kHz. A lower value may sometimes be adequate but the larger value is necessary to provide robust results across a variety of speakers and speaking environments. In z-transform notations, the predictor is represented as

$$P_s(z) = \sum_{k=1}^{p} a_k z^{-k} \tag{1}$$

where the coefficients a_k are called predictor coefficients. In our studies so far, we have used a modified form of the covariance LPC method (together with the correction for missing high frequencies in the signal) to determine the predictor coefficients [5]. The first two steps in this modified procedure are identical to the usual covariance method [11]. Let s_n be the nth speech sample in a block of speech data consisting of $N + p$ samples. In the covariance method, a matrix Φ and a vector c are computed from the speech samples. The element in the ith row and the jth column of the matrix Φ and the ith element of the vector c are given by

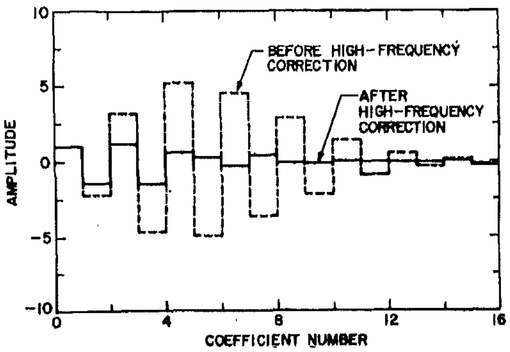
$$\phi_{ij} = \sum_{n=p+1}^{N+p} s_{n-i} s_{n-j}, \qquad i = 1, 2, \cdots, p; \quad j = 1, 2, \cdots, p$$

and

$$c_i = \sum_{n=p+1}^{N+p} s_n s_{n-i}, \qquad i = 1, 2, \cdots, p \tag{2}$$

respectively. The covariance matrix Φ is first expressed as the product of a lower triangular matrix L and its transpose L^t by Cholesky decomposition. Next, a set of linear equations $Lq = c$ is solved. It can then be shown that the partial correlation at delay m is given by

$$r_m = q_m / \epsilon_{m-1}{}^{1/2} \tag{3}$$



BEFORE HIGH-FREQUENCY CORRECTION
AFTER HIGH-FREQUENCY CORRECTION
AMPLITUDE
COEFFICIENT NUMBER
10
5
0
-5
-10
0
4
8
12
16

Fig. 2. Predictor coefficients from LPC analysis of speech before high-frequency correction (broken lines) and after high-frequency correction (solid lines).

where q_m is the mth component of q, and ϵ_m is the mean-squared prediction error at the mth step of prediction. The prediction error is given by

$$\epsilon_m = c_0 - \sum_{i=1}^{m-1} q_i^2 \tag{4}$$

where c_0 is the energy of the speech signal. The partial correlations are transformed to predictor coefficients using the well-known relation between the partial correlations and the predictor coefficients for all-pole filters [12, p. 110]. The modified procedure ensures that all the zeros of the polynomial $1 - P_s(z)$ are inside the unit circle.

The high-frequency correction is necessary due to the gradual rolloff in the amplitude-versus-frequency response of the low-pass filter used in analog-to-digital conversion of the speech signal. The missing high-frequency components in the sampled speech signal near half the sampling frequency produce artificially low eigenvalues of the covariance matrix Φ corresponding to eigenvectors related to such components. These small eigenvalues produce in turn artificially high values of the predictor coefficients after matrix inversion. An example of the predictor coefficients obtained without any high-frequency correction is shown in Fig. 2 (broken line). The covariance matrix of the low-pass filtered speech is almost singular, thereby resulting in a practically *nonunique* solution of the predictor coefficients. Thus, a variety of different predictor coefficients can approximate the speech spectrum equally well in the passband of the low-pass filter. These large predictor coefficients, if used directly for prediction in the coder of Fig. 1, create difficulties. Note that the predictor P, although derived from the original speech signal, is used to predict on the basis of *coded* speech samples which contain an appreciable amount of quantizing noise near half the sampling frequency. The quantizing noise components in the difference signal q_n then can become so large so as to swamp the prediction residual of the speech signal. These problems can be avoided by artificially filling in the missing high frequencies in the digitized speech signal. A procedure for providing this high-frequency correction is described in [5]. The predictor coefficients obtained after high-frequency correction are shown by a solid line in Fig. 2. The power gain (sum of the squares of the predictor coefficients) for successive time frames in a speech utterance

Exhibit T Page 401

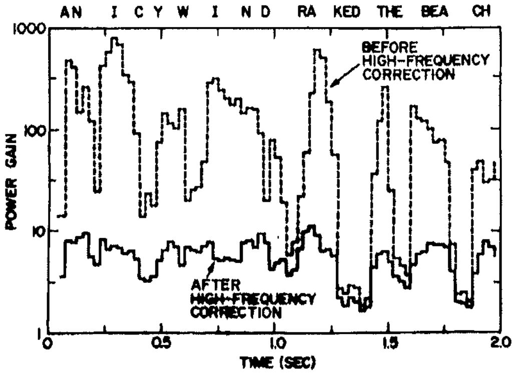

AN I C Y W I N D RA KED THE BEA CH
BEFORE HIGH-FREQUENCY CORRECTION
AFTER HIGH-FREQUENCY CORRECTION
POWER GAIN
TIME (SEC)

Fig. 3. Sum of the squares of the predictor coefficients (power gain) for consecutive time frames in a speech utterance before high-frequency correction (broken line) and after high-frequency correction (solid line). The speech utterance, "An icy wind raked the beach," was spoken by a male speaker.

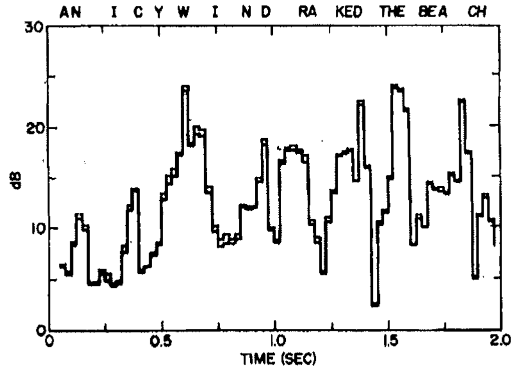


Fig. 4. Prediction gain for consecutive time frames in a speech utterance before high-frequency correction (broken line) and after high-frequency correction (solid line). The speech utterance was the same one used in Fig. 3.

before (broken line) and after (solid line) high-frequency correction is shown in Fig. 3. Although the predictor coefficients with and without high-frequency correction are very different, the prediction errors for the two cases are almost identical. Fig. 4 shows the prediction gain (expressed in decibels) before and after high-frequency correction for the same utterance as was used in Fig. 3. The prediction gain was determined from the LPC spectrum over a frequency range from 0 to 3000 Hz. Speech spectra computed from the two sets (see Fig. 2) of predictor coefficients are shown in Fig. 5. The two spectra are very similar and differ appreciably only in the region near 4 kHz (half the sampling frequency).

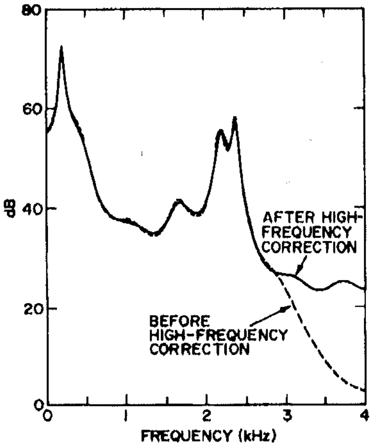

AFTER HIGH-FREQUENCY CORRECTION
BEFORE HIGH-FREQUENCY CORRECTION
dB
FREQUENCY (kHz)

Fig. 5. Spectral envelopes of speech based on LPC analysis before high-frequency correction (solid curve) and after high-frequency correction (dotted curve).

Prediction Based on Spectral Fine Structure

Adjacent pitch periods in voiced speech show considerable similarity. The quasi-periodic nature of the signal is present—although to a lesser extent—in the difference signal obtained after prediction based on spectral envelope. The periodicity of the difference signal can be removed by further prediction. Let the nth sample of the difference signal after the spectral prediction be given by

$$d_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} \tag{5}$$

where s_n is the nth sample of the speech signal, and a_k is the kth predictor coefficient as defined in (1). The predictor for the difference signal can be represented in the z-transform notations by

$$P_d(z) = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1}. \tag{6}$$

The delay M of the predictor $P_d(z)$ is defined as the delay for which the normalized correlation coefficient between d_n and d_{n-M} is highest. The value of M is the equivalent in number of samples of a relatively long delay in the range 2-20 ms. In most cases, that is, when the signal is periodic, this delay would correspond to a pitch period (or possibly, an integral number of pitch periods). The delay M would be random for nonperiodic signals. The coefficients β_1, β_2, and β_3 are determined by minimizing the mean-squared difference between d_n and its predicted value based on the decoded samples. The minimization procedure leads to a set of simultaneous *linear* equations in the three unknowns β_1, β_2, and β_3.

The high-frequency components of the difference signal frequently show less periodicity as compared to the low-frequency components. The three amplitude coefficients β_1, β_2, and β_3 provide a frequency-dependent gain factor in the pitch prediction loop. Moreover, due to a fixed sampling frequency unrelated to pitch period, individual samples of the difference signal do not show a high period-to-period correlation. The third-order pitch predictor provides an interpolated value with much higher correlation than the individual samples.

Combining the Two Types of Adaptive Prediction

The two types of prediction can be combined serially in either order to produce a combined predictor. The order in which the two predictors are combined is important for time-varying predictors. In the earlier work on adaptive predictive coders for speech signals [4], the first predictor was based on the spectral fine structure (pitch). The prediction residual after pitch prediction was used to determine the coefficients of a short-delay predictor with six coefficients.

Exhibit 1 Page 402

Recent studies on APC suggest that it is preferable to use the short-delay predictor (based on spectrum envelope of the speech signal) first [5]. The combined predictor is expressed in the z-transform notation as (short-delay predictor first)

$$P(z) = P_s(z) + P_d(z)[1 - P_s(z)] \tag{7}$$

where $P_s(z)$ and $P_d(z)$ are the two predictors based on spectral envelope and fine structure, respectively. The combined predictor for the case when the pitch predictor is used first is expressed as

$$P(z) = P_d(z) + P_s(z)[1 - P_d(z)]. \tag{8}$$

The relative ordering of the two kinds of prediction (i.e., whether the short-delay predictor is used first or the long-delay predictor is used first) produces coders with very different properties. There are two reasons for this difference. First, the two predictors are very different depending on the order in which the prediction is done. Second, the predictors being time varying, the order of the two predictors cannot be changed without influencing the prediction characteristics of the combined predictor.

Examples of the difference signals after each stage of prediction together with the original speech signal are illustrated in Fig. 6. The difference signal after the first prediction based on the spectral envelope is amplified by 10 dB in the display and that after the pitch prediction is amplified by an additional 10 dB. The prediction residual after two stages of prediction is quite noise-like in nature. Its spectrum—including both envelope and fine structure—is approximately white during steady speech segments. This, however, is not the case during fast transitional segments. The first-order probability density function of the prediction residual samples (after *both* spectral and pitch prediction) is nearly Gaussian. Fig. 7 shows a typical example obtained from a speech utterance approximately 2 s in duration.

Signal-to-Noise Ratio Improvement

The adaptive predictive coder of Fig. 1 provides an improvement in the signal-to-noise ratio (SNR) over a PCM coder using the same quantizer. The improvement is realized because the power of the quantizer input signal q_n is much smaller than that of the input speech signal. The maximum possible gain in the SNR is generally assumed to be equal to the prediction gain defined as the ratio of the power in the speech signal to the power in the prediction residual signal obtained by predicting an input speech sample from previous input speech samples. This is strictly true only if the quantization noise power at every frequency is less than the signal power at that frequency. The predictor P in Fig. 1 is used to predict the current value of the input speech signal based on the previous *reconstructed* speech samples. Each reconstructed speech sample is the sum of the input speech sample and the noise added by the quantizer. The quantizing noise contributes additional power at the input to the quantizer, thereby decreasing the gain in the SNR. The rate distortion theory allows one to calculate the *theoretical* maximum possible improvement in SNR for Gaussian signals [13]. Fig. 8 compares the

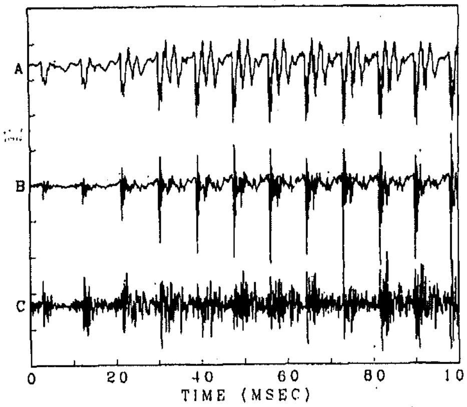


Fig. 6. (A) Speech waveform. (B) Difference signal after prediction based on spectral envelope (amplified 10 dB relative to the speech waveform). (C) Difference signal after prediction based on pitch periodicity (amplified 20 dB relative to the speech waveform).

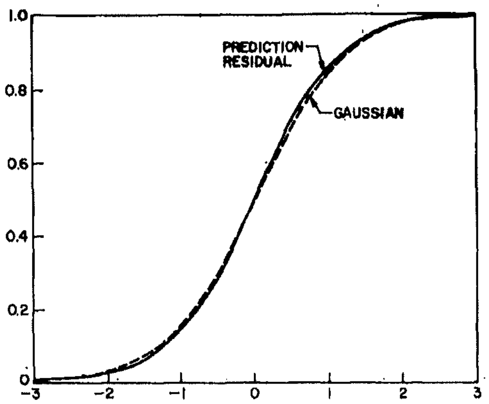


Fig. 7. First-order cumulative amplitude distribution function for the prediction residual samples (solid curve). The corresponding Gaussian distribution function with the same mean and variance is shown by the dashed curve.

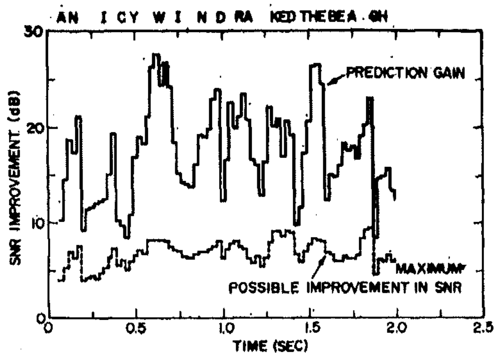


Fig. 8. Prediction gain and the highest possible improvement in SNR according to the rate distortion theory for consecutive time frames in a speech utterance.

Exhibit T Page 403

highest possible improvement in SNR according to rate distortion theory with the prediction gain. The spectra for different time frames were obtained by LPC analysis of a sentence-length speech utterance. The quantizer was assumed to have an SNR of 10 dB. As can be seen, the maximum possible improvement in SNR is considerably smaller than the prediction gain.

Encoding of Predictor Parameters

The digital channel in an adaptive predictive coding system must carry information about the parameters of the time-varying filter at the receiver. Efficient coding of the parameters is necessary to keep the total bit rate to a minimum.

The block diagram of the receiver of the APC system is shown in Fig. 9. It consists of two linear filters each with a predictor in its feedback loop. The first feedback loop includes the long-delay (pitch) predictor $P_d(z)$ which restores the pitch periodicity of voiced speech. The second feedback loop which includes the short-delay predictor $P_s(z)$ restores the spectral envelope.

Direct quantization of the predictor coefficients a_k is not recommended [14]-[16]. It is preferable to quantize and encode the partial correlation coefficients; either a quantizer with nonuniformly spaced levels must be used or the partial correlations must be suitably transformed to make their probability density functions more uniform. Two kinds of transformations, inverse sine and inverse hyperbolic tangent, have been used for this purpose. The precision with which each partial correlation must be encoded varies from one coefficient to another. In general, the higher order coefficients need less precision than the lower order coefficients.

We have found the uniform quantization of the inverse sine of the partial correlations to be a reasonable solution to the quantization problem. The range of variation of the partial correlations was found by computing the probability density function for each coefficient; the minimum and maximum values were selected to include 99.6 percent of the entire range of their variations. The number of quantization levels for the different coefficients at any desired bit rate was determined by using an iterative procedure in which the distribution of bits was varied to minimize the spectral error (mean-squared error in the logarithmic spectrum). Starting from a uniform distribution of bits for the different coefficients, two partial correlation coefficients were identified, one which was most effective in reducing the spectral error by increasing the number of bits and another which was least effective. The bit assignment was increased for the most effective coefficient and was decreased by the same amount for the least effective coefficient, thus keeping the total number of bits to be constant. Any coefficient, which had only one bit assigned to it, was not considered for further reduction in its allocation of bits. Table I shows the distribution of bits for the first 20 coefficients, using a total of 50 bits. The range of each partial correlation (after inverse sine conversion) is also shown in the table. These results are based on a total of approximately 60 s of speech spoken by both male and female speakers. The speech was low-pass filtered to 3.6 kHz and sampled at 8 kHz. The high frequencies in the sampled speech signal were preemphasized using a filter with the transfer function $1 - 0.4z^{-1}$. Our experience with high-frequency preemphasis has been that it is preferable to use only a mild degree of emphasis. This is in contrast to the common practice of using a filter $1 - z^{-1}$ for emphasizing high frequencies. Such a strong emphasis of high frequencies creates difficulties at the receiver. Although the spectral balance of the speech signal can be restored by using a proper inverse filter, the low-frequency components of the quantizer noise are greatly magnified by the inverse filter. The mild emphasis limits this increase of low frequencies at the receiver to relatively small amplitudes.

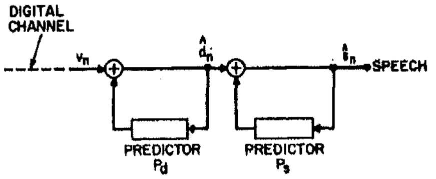


Fig. 9. Block diagram of the receiver of the APC system.

TABLE I
BIT ALLOCATION AND RANGE OF VARIATION OF INVERSE SINE OF PARTIAL CORRELATIONS

Coefficient	Minimum	Maximum	Bits
1	−1.380	0.540	5
2	−0.300	1.230	5
3	−1.110	0.390	4
4	−0.300	1.110	4
5	−0.900	0.450	3
6	−0.300	0.900	3
7	−0.750	0.570	3
8	−0.450	0.600	3
9	−0.540	0.450	3
10	−0.450	0.360	3
11	−0.390	0.390	2
12	−0.330	0.330	2
13	−0.240	0.270	2
14	−0.270	0.300	2
15	−0.210	0.300	1
16	−0.240	0.330	1
17	−0.180	0.300	1
18	−0.210	0.300	1
19	−0.180	0.270	1
20	−0.180	0.270	1

Informal listening test show that the bit assignment shown in Table I does not produce any significant additional distortion in the reproduced speech signal as a result of quantization of partial correlations. The distortion is still small although audible when a total of 40 bits are used for encoding 20 partial correlations. Furthermore, it is generally sufficient to reset the coefficients of the short-delay predictor both at the transmitter and the receiver once every 10 ms. The distortion is not increased significantly even when the coefficients are reset once every 20 ms. The total bit rate for the coefficients depends both on the number of coefficients and the time intervals at which a new set of coefficients are determined. For example, a bit rate of 4600 bits/s is realized by using 16 coefficients reset to their new values every 10 ms. The bit rate is reduced to 2300 bits/s if the time interval for resetting the coefficients is changed to 20 ms.

The delay parameter M of the long-delay (pitch) predictor P_d needs approximately 7 bits of quantization accuracy. It is

Exhibit T Page 404

TABLE II
BIT ALLOCATION AND RANGES FOR THE PITCH PREDICTOR PARAMETERS

Parameter	Minimum	Maximum	Bits
M	20	147	7
b_1	−1.2	1.2	5
b_2	−1.0	1.0	4
b_3	−1.0	1.0	4

desirable to transform the three amplitude coefficients β_1, β_2, and β_3 of the pitch predictor prior to quantization as shown below.

$$b_1 = \log(\beta_1 + \beta_2 + \beta_3)$$

$$b_2 = \beta_1 - \beta_3$$

and

$$b_3 = \beta_1 + \beta_3. \tag{9}$$

The bit assignment and the ranges of the transformed parameters b_1, b_2, and b_3 are shown on Table II. The pitch predictor must be reset once every 10 ms to be effective resulting in a bit rate of 2000 bits/s for the pitch predictor parameters.

The total bit rate for the parameters of the two predictors is 4300 bits/s if the coefficients of the short-delay predictor are reset every 20 ms and 6600 bits/s if they are reset every 10 ms. The rms value of the prediction residual needs about 6 bits of quantization and must be reset once every 10 ms. The side information thus needs somewhere between 4900 and 7200 bits/s.

III. GENERALIZED PREDICTIVE CODER WITH NOISE SHAPING

Traditionally, waveform coders have attempted to minimize the rms difference between the original and coded speech waveforms. However, it is now well recognized that the subjective perception of the signal distortion is not based on the rms difference (error) alone. In designing a coder for speech signals, it is necessary to consider both the short-time spectrum of the quantizing noise and its relation to the short-time spectrum of the speech signal. Due to auditory masking, the noise in the formant regions (frequency regions where speech energy is concentrated) is masked to a large degree by the speech signal itself. Thus, the frequency components in the noise around the formant regions can be allowed to have higher energy relative to the components in the interformant regions. Similar tradeoffs can be realized between low and high frequency regions.

A simple method of providing flexibility in controlling the spectrum of the quantizing noise is to use a conventional APC system with a prefilter and a postfilter [17]. Such a system (called D*PCM by Noll [17]) is shown in Fig. 10. The speech signal s_n is prefiltered by a time-varying filter $1 - R$ to generate a new signal y_n. The predictor P_A is optimized for predicting the signal y_n. It includes two predictors: a predictor P_y based on the spectral envelope of the signal y_n and another predictor P_d based on the spectral fine structure. The combined predictor is given by

$$P_A(z) = P_y(z) + P_d(z)[1 - P_y(z)]. \tag{10}$$

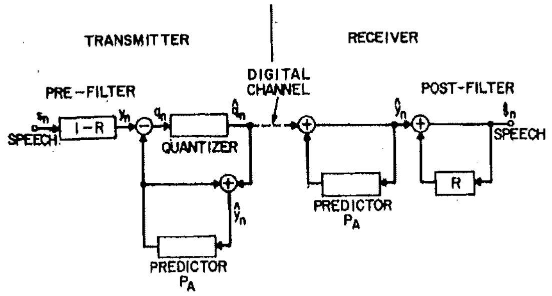


Fig. 10. Block diagram of a generalized predictive coder using pre- and postfiltering to achieve desired spectrum of the quantizing noise. The desired noise spectrum is realized by proper selection of the filter $1 - R$.

The noise-shaping properties of this configuration are described more conveniently in the frequency domain. Let $S(\omega)$ and $\hat{S}(\omega)$ represent the Fourier transforms of the input and the output speech signals, respectively. Similarly, let $Q(\omega)$ and $\hat{Q}(\omega)$ represent the Fourier transforms of the quantizer input and output signals, respectively. One can then write[1]

$$\hat{S}(\omega) - S(\omega) = [1 - R(\omega)]^{-1}[\hat{Q}(\omega) - Q(\omega)] \tag{11}$$

where $1 - R(\omega)$ is the Fourier transform of the prefilter's impulse response. Under the assumption that the quantizer noise is white (true only for quantizers which have a fairly large number of levels covering the entire range of signal at the quantizer input), the spectrum of quantizing noise in the reconstructed speech signal is given by

$$N(\omega) = \sigma_q^2 \,|[1 - R(\omega)]|^{-2} \tag{12}$$

where σ_q^2 is the variance of the quantizer noise appearing at the output of the quantizer. With fine quantization, any desired spectrum of the noise can be achieved by appropriate selection of the filter R.

A different but functionally equivalent configuration for the noise-shaping coder has been suggested by Kimme and Kuo [18]. Fig. 11 illustrates how the shaping of the quantization noise spectrum is achieved. The quantization noise (difference between the output and the input of the quantizer) is filtered by a linear filter with frequency response $F_B(\omega)$ and is subtracted from a prediction residual signal. The resulting difference signal then forms the input to the quantizer. The Fourier transform of the quantizer noise appearing in the reconstructed speech signal can be expressed as

$$\hat{S}(\omega) - S(\omega) = [1 - P_B(\omega)]^{-1}[1 - F_B(\omega)][\hat{Q}(\omega) - Q(\omega)] \tag{13}$$

where the upper case letters again represent various variables in the Fourier transform domain. The spectrum of the quantiz-

[1] Equation (10) is strictly true only when both the predictors and the noise-shaping filters are time-invariant. For speech signals, one is tempted to replace the infinite-time transforms by short-time transforms. This procedure is approximately valid, provided the impulse response of each filter lasts only over time intervals during which the filter response does not change appreciably. This is true usually for the prediction or filtering based on the short-time spectral envelope, but not for the pitch predictor. The impulse response of the pitch predictor typically lasts over several pitch periods for voiced sounds.

Exhibit T Page 405

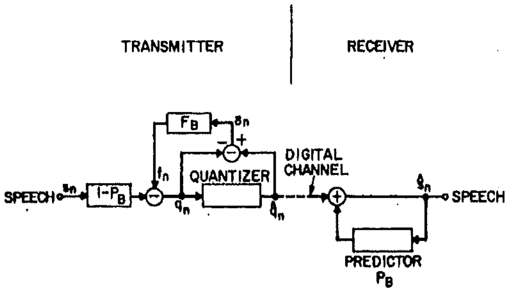


Fig. 11. Block diagram of another generalization of a predictive coder with adjustable noise spectrum. The desired spectrum is achieved by adjusting the noise-feedback filter F_B.

ing noise in the reconstructed speech signal is given by

$$N(\omega) = \sigma_q^2 |[1 - F_B(\omega)][1 - P_B(\omega)]^{-1}|^2. \tag{14}$$

The two coders shown in Figs. 10 and 11 are equivalent if

$$F_B(\omega) = P_A(\omega)$$

and

$$1 - P_B(\omega) = [1 - P_A(\omega)]\,[1 - R(\omega)]. \tag{15}$$

It must be recognized that the predictor P_B in Fig. 11 is the predictor for the speech signal. The predictor P_A in Fig. 10, on the other hand, is the predictor for the prefiltered speech signal.

Yet another different but functionally equivalent configuration for a noise-shaping coder has been proposed by Makhoul [9]. This particular configuration uses a somewhat different form of noise feedback and is illustrated in Fig. 12. With fine quantization, the spectrum of the quantizing noise in the reconstructed speech signal is given by

$$N(\omega) = \sigma_q^2 |[1 - F_C(\omega)]|^2. \tag{16}$$

This generalization of the predictive coder is also equivalent to the coder of Fig. 10 if

$$1 - P_C(\omega) = [1 - P_A(\omega)]\,[1 - R(\omega)]$$

and

$$1 - F_C(\omega) = [1 - R(\omega)]^{-1}. \tag{17}$$

The different noise-shaping generalizations of predictive coders shown in Figs. 10–12 are functionally equivalent and differ merely in the manner in which the predictors and the noise-shaping filters are configured. In practice, they provide nearly identical performance. The selection of a particular configuration depends primarily on the desired shape of the noise spectrum. For example, the noise-shaping configuration of Fig. 10 is clearly preferable if the noise spectrum includes only poles or zeros. Similarly, the configuration shown in Fig. 11 is preferable if the noise spectrum is a pole-zero spectrum with the same set of poles as the speech spectrum has.

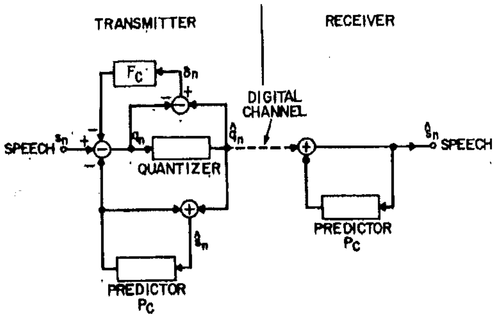


Fig. 12. Block diagram of yet another generalization of a predictive coder for controlling the spectrum of quantizing noise in output speech.

Perceptual Criteria for Selecting the Noise-Shaping Filter

Since the various generalizations of APC to provide noise shaping are equivalent, we will limit the discussion to the coder shown in Fig. 10. The filter $1 - R$ in Fig. 10 provides flexible control of the noise spectrum and can be chosen to minimize an error measure in which the noise is weighted according to some subjectively meaningful criterion. For example, an effective error measure can be defined by weighting the noise power at each radian frequency ω by a function $W(\omega)$. For a fixed quantizer, the spectrum of the output noise is proportional to $G(\omega) = |[1 - R(\omega)]|^{-2}$. One could choose R to minimize

$$E_n = \int_0^{\pi} G(\omega) W(\omega)\, d\omega \tag{18}$$

under the constraint [5]

$$\int_0^{\pi} \log G(\omega)\, d\omega = 0. \tag{19}$$

The minimum is achieved if

$$\log G(\omega) = -\log W(\omega) + \frac{1}{\pi} \int_0^{\pi} \log W(\omega)\, d\omega. \tag{20}$$

The function $[1 - R]^{-1}$ is the minimum-phase transfer function with spectrum $G(\omega)$ and can be obtained by direct Fourier transformation or spectral factorization. A particularly simple solution to this problem is obtained by transforming $G(\omega)$ to an autocorrelation function by Fourier transformation. By using a procedure similar to LPC analysis, the autocorrelation function can be used to determine a set of predictor coefficients. The predictor coefficients determined in this manner are indeed the desired filter coefficients for the filter R. The solution is considerably simplified if the noise-weighting function $W(\omega)$ is expressed in terms of a filter transfer function whose poles and zeros lie inside the unit circle.

A better procedure for achieving optimal subjective performance is to choose the filter R such that the subjective loudness (or audibility) of quantizing noise in the presence of the speech signal is minimized. The loudness of quantization noise depends on the excitation patterns of *both* the speech

Exhibit T Page 406

signal and the quantization noise along the basilar membrane in the inner ear. Due to the nonstationary nature of the speech signal, its short-time spectrum and, therefore, its excitation pattern along the basilar membrane vary continuously with time. The detailed procedure for computing the time-varying loudness of speech signals and noise is described in [7] and [8]. An efficient procedure for designing the optimal noise-shaping filter to minimize the subjective loudness of the quantizing noise is described in [19].

Several arbitrary but illustrative choices for the filter $1-R$ can be considered. The first obvious choice is to set $R = 0$. The spectrum of output noise is white with fine quantization, producing a very high SNR in the formant regions, but a poor one in between the formants where the magnitude of the signal spectrum is low. A very high SNR at any frequency, higher than what is necessary to produce zero loudness, is a waste. Since the noise shaping can only redistribute noise power from one frequency to another as provided in (19), it is better to increase the noise in the formant regions to a level where it is barely perceptible and to use this increase to reduce the noise in between formant regions. Thus, $R = 0$ is not a good choice. At the other extreme, one can set the quantizing noise spectrum to be proportional to the speech spectrum, i.e., $R = P_s$. This would be a good choice if our ears were equally sensitive to quantizing distortion at all frequencies. However, this is not so. An intermediate choice is given by

$$1-R(z) = \left[1-\sum_{k=1}^{p} a_k z^{-k}\right] \Big/ \left[1-\sum_{k=1}^{p} a_k \alpha^k z^{-k}\right] \qquad (21)$$

where α is an additional parameter controlling the increase in the noise power in the formant regions. The filter R changes from $R = 0$, for $\alpha = 1$, to $R = P_s$, for $\alpha = 0$. At a sampling frequency of 8 kHz, α is typically 0.73. An example of the envelope of the output quantizing noise spectrum together with the corresponding speech spectrum is shown in Fig. 13.

IV. QUANTIZATION OF PREDICTION RESIDUAL AT LOW BIT RATES

The digital channel in an APC system carries two separate kinds of information: one about the quantized prediction residual and the other about the time-varying predictors and the step size of the quantizer. The transmission of the prediction residual usually requires a significantly larger number of bits per second in comparison to the side information.[2] Efficient quantization of the prediction residual is thus essential in achieving the lowest possible bit rate for a given speech quality.

[2] As discussed in Section II, the transmission of side information requires approximately 4 kbits/s. The bit rate for the prediction residual depends both on the sampling frequency and the number of levels used for quantizing the prediction residual. As an example, this bit rate would be 12.8 kbits/s for a three-level quantizer at a sampling frequency of 8 kHz.

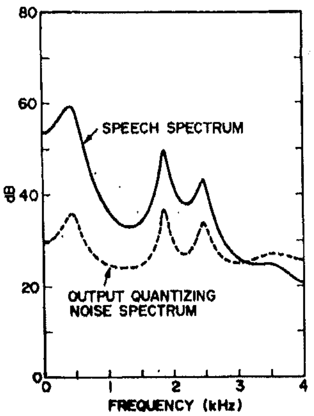


Fig. 13. Spectral envelopes of output quantizing noise (dotted curve) shaped to reduce the perceived distortion and the corresponding speech spectrum (solid curve).

The number of quantization levels must be an integer number. The bit rate for the quantized prediction residual can thus take only a certain discrete set of values. For example, at a sampling frequency of 8 kHz, the bit rate for the prediction residual is 12.8 kbits/s for a three-level quantizer. The next lower bit rate is 8 kbits/s. Since the minimum possible number of levels is two, the bit rate cannot be lower than 8 kbits/s at a sampling frequency of 8 kHz. Moreover, such a coarse quantization, with only two levels per sample, is usually the major source of audible distortion in the reconstructed speech signal. With only two levels, it is difficult to avoid both peak clipping of the prediction residual and the granular distortion due to a finite number of levels in the quantizer. Peak clipping of the prediction residual produces a distortion which is, in many respects, similar to the slope-overload distortion in delta modulators [20]. In addition, peak clipping can produce occasional "pops" and "clicks" in the speech signal. A large step size chosen to avoid peak clipping introduces a significant amount of granular (random) noise similar to the one encountered in PCM systems with coarse quantization. An example of peak clipping in a two-level quantizer is illustrated in Fig. 14. The figure shows (a) the prediction residual, (b) the quantizer input, (c) the quantizer output, (d) the reconstructed difference signal $\hat{d}_n$, (e) the original difference signal d_n, (f) the reconstructed preemphasized speech signal $\hat{s}_n$, and (g) the original preemphasized speech signal s_n. The difference signals d_n and $\hat{d}_n$ are amplified by 6 dB relative to the preemphasized speech signal s_n in the illustration. The prediction residual, the quantizer input, and the quantizer output are further amplified by 6 dB. The peak clipping of the prediction residual is evident near the beginning of the pitch periods in both the reconstructed difference and speech signals. We find, in general, that amplitude variations in the quantizer input are often large and cannot be handled properly by a two-level quantizer.

Improved Quantization at Low Bit Rates

Recent studies [21] indicate that accurate quantization of high-amplitude portions of the prediction residual is necessary for achieving low perceptual distortion in the decoded speech

Exhibit T Page 407

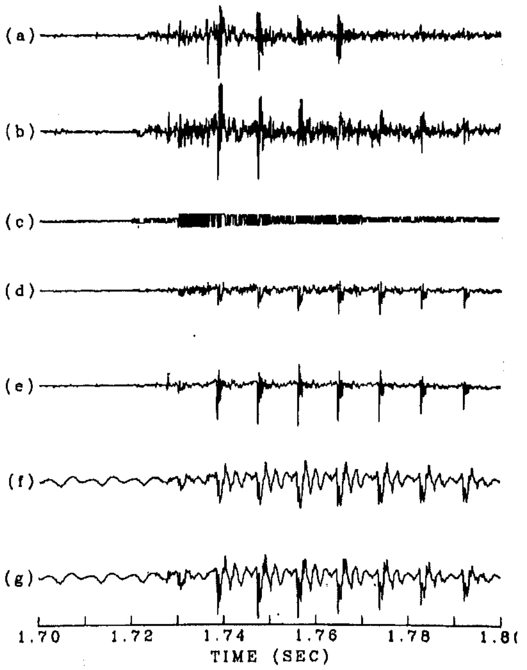


Fig. 14. Waveforms of different signals in a predictive coder with two-level quantizer: (a) the prediction residual, (b) the quantizer input, (c) the quantizer output, (d) the reconstructed difference signal $\hat{d}_n$, (e) the original difference signal d_n, (f) the reconstructed preemphasized speech signal $\hat{s}_n$, and (g) the original preemphasized speech signal s_n. The difference signals d_n and $\hat{d}_n$ are amplified by 6 dB relative to the preemphasized speech signal. The prediction residual, the quantizer input, and the quantizer output are amplified by 12 dB relative to the speech signal.

signal. Moreover, very little or no distortion is audible in the presence of severe center clipping of the prediction residual. This implies that quantization of each sample of the prediction residual with the same step size is not a good procedure. It is better to use most of the available bits for encoding the high-amplitude portions of the prediction residual. To keep the bit rates within a specified value, the prediction residual can be severely center clipped prior to quantization by a multilevel quantizer. The center clipping produces a large number of zeros at the output of the quantizer. The entropy of the quantized signal is thus quite low, although the high-amplitude portions of the prediction residual are quantized into many levels. A block diagram illustrating this improved quantization procedure is shown in Fig. 15.

Maximum Number of Quantizer Levels

We will consider only forward-adaptive quantizers in which the step size is reset to a new value at the beginning of each frame and is held constant for the entire frame. This step size is transmitted to the receiver as part of the side information. It is of considerable importance to know the minimum number of quantizer levels necessary to produce speech at the receiver with no perceptual distortion. For this test, the speech signal was sampled at 8 kHz, the predictor P_s had 16 taps, the predictor P_d had three taps, and the noise-shaping filter $1 - R$ was chosen according to (21) with $\alpha = 0.73$. Our results indicate that 15 levels, distributed uniformly to cover the entire range of prediction residual amplitudes (after both types of prediction) in a frame, are sufficient. The quantizer step size is chosen to be $2 \times V_{\text{peak}}/(n_q - 1)$ where V_{peak} is the peak value (maximum absolute value) of the prediction residual samples in a frame and n_q is the number of quantizer levels. A step size chosen on the assumption of the Gaussian distribution for the amplitude of the prediction residual produced significant peak clipping in the quantization process. The peak value of the quantizer input is usually greater than V_{peak}, and therefore it is difficult to avoid peak clipping completely.

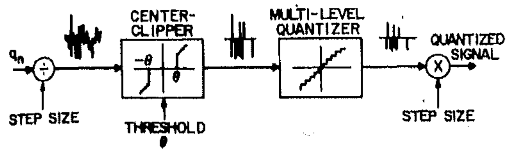


Fig. 15. Improved procedure for quantizing prediction residual using a center-clipping quantizer.

Selection of Center-Clipping Threshold

We investigated several methods for adjusting the threshold of center clipping. These included methods in which the threshold was set once for each frame as well as methods in which the threshold was adjusted at each sampling instant. As an example, we set the threshold in each frame to be proportional to the rms value of the prediction residual v_n in that frame. The first-order entropy of quantized prediction residual, averaged over sentence-length utterances, decreased with the increase in the threshold of center clipping as shown in Fig. 16. The center-clipping threshold can thus be used to obtain any desired value of the entropy. The speech quality is still fairly good for threshold values up to twice the rms value! In this case, the number of nonzero samples in the quantized prediction residual is only 10 percent of the total. However, there is considerable variation in the number of nonzero samples, and therefore in the entropy, from one frame to another.

Another choice for adjusting the center-clipping threshold would be to select a value which will produce a fixed number of nonzero samples at the quantizer output in each frame. However, it is not possible to select such a threshold value in advance at the beginning of the frame. Our experience has been that it is necessary to vary the center-clipping threshold on a sample-by-sample basis in order to avoid large variation in the number of nonzero samples. A somewhat arbitrary but still reasonable procedure for adjusting the threshold is given below.

Let γ be the desired fraction of the quantized samples which are nonzero. A typical value of γ is 0.10. Let θ_0 be an initial estimate of the center-clipping threshold. The threshold at the nth sampling instant is given by

$$\theta_n = \theta_0 [(\mu_n/\gamma) |\overline{q_n}| / |\overline{v_n}|]^{1/2} \tag{22}$$

where $|\overline{q_n}|$ is the running average on a sample-by-sample basis of the absolute value of the quantizer input signal q_n, $|\overline{v_n}|$ is the corresponding running average for the signal v_n, and μ_n is the running average for the actual fraction of the quantized samples which are nonzero at the output of the quantizer. The averages are computed by using an integrator with a time constant of 5 ms.

Exhibit T Page 408

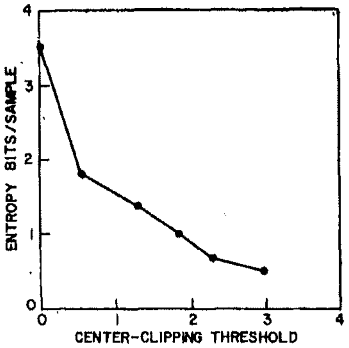


Fig. 16. First-order entropy of the quantizer output as a function of the ratio of the threshold of center clipping to the rms value of the prediction residual.

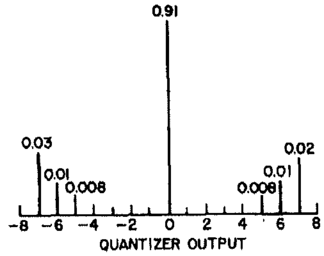


Fig. 17. Distribution of quantizer output levels in the center-clipping quantizer shown in Fig. 15 with the center-clipping threshold adjusted such that only 9 percent of the quantized samples have nonzero values.

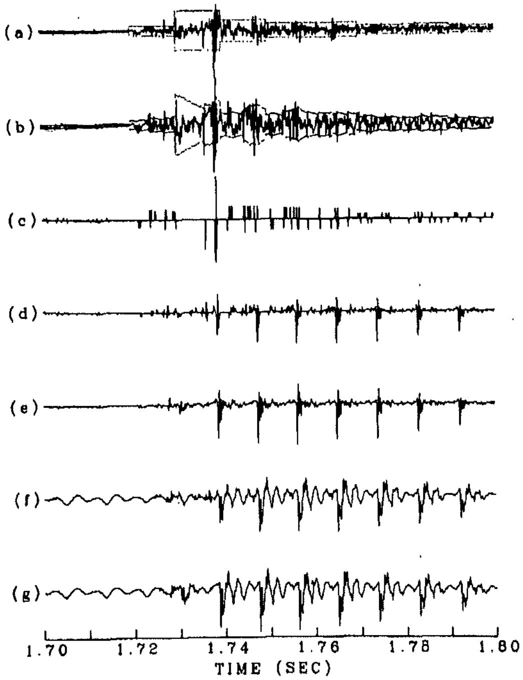


Fig. 18. Waveforms of different signals in a predictive coder with a multilevel quantizer with center clipping shown in Fig. 15. The labels for the different waveforms are identical to the ones shown in Fig. 14. The broken curve in the waveform (a) is the threshold θ_0 selected initially at the beginning of each frame. The broken curve in the waveform (b) is the threshold θ_n adjusted adaptively at each sampling instant.

The center-clipped prediction residual is quantized by a 15-level uniform quantizer with step size = $V_{\text{peak}}/7$. The distribution of quantizer output levels shows that the eight innermost levels, ±1, ±2, ±3, and ±4, remain zero with a very high probability. No additional audible distortion in the speech signal is produced by constraining the quantizer output to have only seven levels, 0, ±5, ±6, and ±7. The number of nonzero samples varies somewhat from one frame to the next but this variation is not excessive. A typical distribution of quantizer output levels is shown in Fig. 17. The probabilities are shown on a logarithmic scale in the figure. The first-order entropy of this distribution is 0.64 bit/sample.

The waveform plots for this improved quantization procedure are shown in Fig. 18. The speech segment is identical to the one shown in Fig. 14, and therefore it is easy to compare the two sets of plots. As before, the figure shows (a) the prediction residual, (b) the quantizer input, (c) the quantizer output, (d) the reconstructed difference signal, (e) the original difference signal, (f) the reconstructed preemphasized speech signal, and (g) the original preemphasized speech signal. The broken curve in the waveform (a) is the threshold θ_0 selected initially at the beginning of each frame. The broken curve in the waveform (b) is the threshold θ_n adjusted adaptively at each sampling instant. It is obvious that the peak clipping of the prediction residual is reduced significantly in comparison to the two-level case shown in Fig. 14. The reduction in peak clipping is achieved without any increase in the step size of the quantizer (indeed, the step size is considerably reduced). Thus, the new quantization procedure produces less peak clipping as well as less granular distortion in comparison to the two-level case. As a consequence, there is considerable improvement in the subjective quality of the reconstructed speech. Informal listening tests with several sentences spoken by both male and female speakers indicate that the reconstructed speech signal has very little or no perceptible distortion. Only in close headphone listening can one detect minute distortion at the beginning of some voiced speech segments. The bit rate for the prediction residual using the center-clipping quantization is only 5.6 kbits/s (0.70 bit/sample × 8000 samples/s). The total bit rate including the side information is approximately 10 kbits/s.

The center-clipping procedure is very effective in reproducing accurately both the waveform and the spectrum of the speech signal at the receiver. Typical examples of the spectra of the original (solid curve) and the reconstructed (dashed curve) speech signals are shown in Fig. 19. The spectra were computed from speech segments 40 ms in duration after applying the Hamming window and the successive speech segments were spaced 20 ms apart. The average SNR was found to be approximately 12 dB. The segmental signal-to-noise ratio for a sentence-length utterance, "An icy wind raked the beach," spoken by a male speaker is shown in Fig. 20 (solid curve). The speech power expressed in dB is also shown in the figure by a dashed curve. The SNR is higher for voiced speech as compared to unvoiced speech, and during voiced portions the SNR is higher for steady segments as compared to transitional segments.

Exhibit 1 Page 409

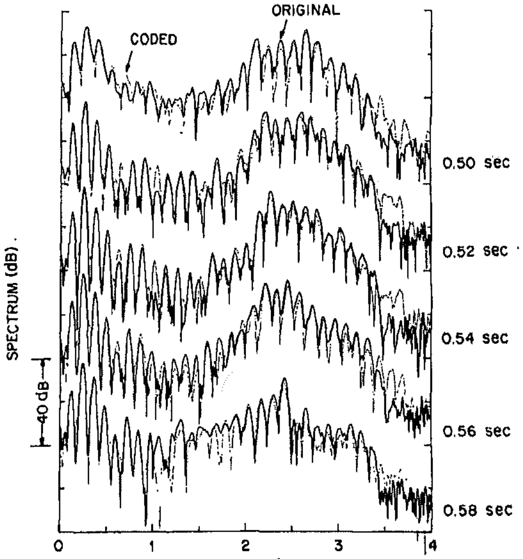


Fig. 19. Examples of spectra of the original speech (solid curve) and the reconstructed speech (dashed curve) waveforms. The spectra were obtained from 40 ms long speech segments using a Hamming window. Successive spectral frames correspond to speech segments 20 ms apart.

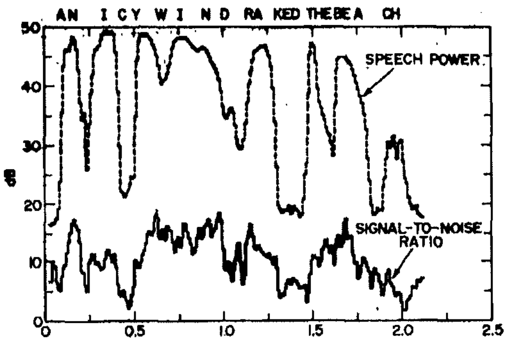


Fig. 20. Segmental signal-to-noise ratio (SNR) for successive time frames for the utterance, "An icy wind raked the beach," spoken by a male speaker. The solid curve is the speech power expressed in dB.

Further experiments with even higher levels of center clipping show that speech quality degrades slowly with increasing number of zeros in the quantizer output. The distortion is small even when the probability of zeros in the quantizer output is increased to 0.95, corresponding to a first-order entropy of 0.45 bit/sample. However, the distortion is quite noticeable when the probability of zeros is increased to 0.98 (first-order entropy = 0.20 bit/sample). The average signal-to-noise ratio in this case is approximately 8 dB. Thus, a reduction in the bit rate of 0.4 bit/sample produces a decrease of 4 dB in the SNR.

Encoding of the Quantized Prediction Residual

The quantized prediction residual needs suitable encoding for efficient transmission over a digital channel. Due to the large number of zeros in the prediction residual (produced by center clipping with a large threshold) it would be highly inefficient to assign the same number of bits to every sample of the prediction residual. Since the entropy of the prediction residual is less than 1 bit/sample, it is necessary to group a number of samples together in a block of appropriate length and use the resulting block of samples as an input symbol for the coder rather than the individual samples. A variable-length code, which assigns shorter codewords for inputs with a higher probability of occurrence and longer codewords for inputs with lower probability of occurrence, can then be used to achieve high coding efficiency. There are two special problems with variable-length codes. First, the digital channels often transmit data at uniform bit rates. One must then provide a buffer between the variable-length codes and the uniform bit rate channel. The center-clipping quantizer makes it particularly easier to manage buffer overflow problems; the threshold of center clipping can be increased or decreased dynamically to control the number of bits going into the buffer and, thus, to prevent overflows. Second, the digital channels often introduce errors in transmitted bits. A variable-length code must be designed to provide a loss of codeword synchronization in the presence of channel errors. One possibility is to use variable-length-to-block codes [22]. These codes use codewords of fixed lengths to represent a variable number of input samples and, thus, have no synchronization problems in the presence of channel errors. The performance of such codes can be made arbitrarily close to the rate-distortion optimum by using a large enough set of codewords.

The variable-length-to-block codes are a generalization of runlength codes and are easy to construct. For example, one could use a code of fixed length to represent a sequence of samples all of zero amplitude terminated by a sample with nonzero amplitude. If the prediction residual is quantized into seven levels 0, ±5, ±6, and ±7, and if the maximum number of zeros in any sequence is limited to 21, then there are at most 127 (6 × 21 + 1) possible sequences, all of which can be represented by a fixed-length 7 bit code. With the probability of zeros in the quantized residual of 0.10, the above code would produce a bit rate of 5.6 kbits/s at a sampling frequency of 8 kHz. For comparison, the first-order entropy of the distribution shown in Fig. 17 is 0.64 bit/sample.

In variable-length-to-block coding, the source sequences of variable length are assigned codewords of constant length. Thus, codeword boundaries can be uniquely decoded even in the presence of channel errors. However, it is still possible for a channel error to cause a codeword to be decoded into a sequence with a different number of samples than what it actually contained. This problem can be resolved by block-to-block coding in which a fixed number of prediction residual samples is coded into sequences containing a fixed number of bits. In one implementation it was found that 240 prediction residual samples (corresponding to a time interval of 30 ms at the sampling frequency of 8 kHz) can be coded into 192 bits (0.80 bit/sample) without introducing any additional distortion in the reproduced speech signal [23].

V. CONTROL OF ERROR SPECTRUM AT LOW BIT RATES

With *fine* quantization (large number of closely spaced quantization levels), the generalized predictive coder of Fig. 10

Exhibit T Page 410

(or any one of its functional equivalents shown in Figs. 11 and 12) is capable of producing any desired shape of the quantizing noise spectrum by proper selection of the filter R. For coarse quantization or with severe center clipping (as described in Section IV) the spectrum of the quantizer error signal is not necessarily white and becomes an important factor in determining the spectrum of the quantization noise at the output of the coder. Fig. 21 shows a typical example (dashed curve) of the spectrum of the quantizing noise appearing in the reconstructed speech signal in the coder described in Section IV. The coder uses a center-clipping quantizer to reduce the bit rate of the prediction residual to an average value of 0.70 bit/sample. For comparison, the spectrum of the input speech signal is also shown (solid curve) in the figure. The noise shaping is done using the prefilter given by (21). As expected, the peaks in the spectral envelope of the quantizing noise occur in the formant regions. However, the fine structure of the quantizing noise spectrum shows considerable pitch periodicity[3] which is not specified in (21). Thus, with coarse quantization a predictive coding system does not provide the required flexibility in adjusting the spectrum of the quantization noise to the desired shape. This is a major shortcoming of a predictive coder with instantaneous quantization. We consider noise shaping to be of crucial importance for realizing very low bit rates. We will discuss in this section a class of coders which represent a further generalization of predictive coders. These coders not only allow one to realize the precise optimum noise spectrum, but also represent the important first step in bridging the gap between waveform coders and vocoders without suffering from their limitations.

Ideally in speech coding, one is interested in finding a sequence of binary digits which after decoding produce a synthetic speech signal which is close to a given speech signal according to a particular fidelity criterion. Speech signals are produced as a result of acoustical excitation of the vocal tract. The filtering action of the vocal tract can then be reproduced at the receiver by a linear filter. Furthermore, the periodic nature of the vocal excitation can also be produced by a linear filter. Thus, a suitable decoder for speech is a *time-varying* linear filter whose parameters are determined by appropriate analysis of the speech signal in a manner similar to one described earlier in selecting a predictor for speech signals. The speech coding problem is then reduced to finding an input sequence at a given bit rate which after decoding produces minimum error according to the particular fidelity criterion. The above ideas are illustrated in the block diagram of Fig. 22. The input sequence v_n is filtered by a known time-varying filter H to produce a synthetic speech sequence $\hat{s}_n$. The synthetic sequence $\hat{s}_n$ is compared with the original speech sequence s_n and the resulting difference signal is modified by the weighting network W to produce a perceptually weighted

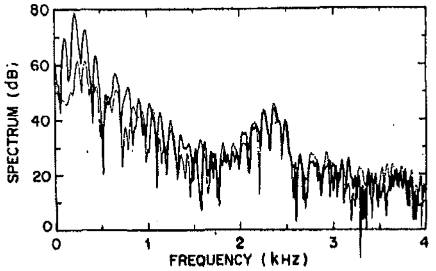


Fig. 21. Example of the spectrum of the quantizing noise (dashed curve). The spectrum for the corresponding speech signal is illustrated by the solid curve. The spectra were obtained from 40 ms long speech segments using a Hamming window.

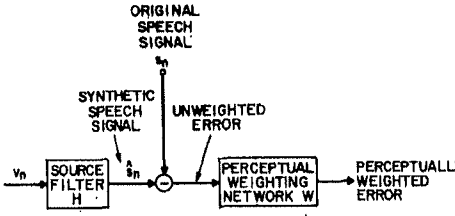


Fig. 22. Block diagram of a speech coder suitable for obtaining a precise noise spectrum even at low bit rates.

error signal. The object of the encoder is to generate the optimum input sequence at a given bit rate to minimize the energy in the weighted error signal (averaged over time intervals approximately 10-20 ms in duration). The weighting network W represents our knowledge of how human perception treats the differences between the original and the synthetic speech signals.[4] The weighting network can be designed so that the loudness of noise in the synthetic speech signal is minimized. It is easy to see that W plays the same role as the prefilter $1 - R$ did in the predictive coder [10]. Indeed, $W = 1 - R$. The procedure for determining the optimum W is then identical to the method described earlier in Section III for determining the filter $1 - R$.

There are several ways in which one could impose the constraint that the input sequence v_n has a specified bit rate. One possibility is to use tree codes [24]. These codes have been shown to perform arbitrarily close to the rate-distortion bound for memoryless sources. Although there has been considerable interest recently in tree coding of speech signals, much of this work has not focused on the noise shaping problem [25]-[27]. Recent work of Wilson and Husain has addressed this problem, but restricted to a *fixed* frequency-weighted error criterion [28]. It is essential for achieving optimum performance at low bit rates that both the source filter H and the error-weighting filter W be adaptive. The vocal tract cannot be represented by a *fixed* linear filter in any useful manner. Similarly, the perception of error in the synthetic speech signal cannot be represented by a *fixed* error spectrum.

[3] It could be argued that the presence of pitch periodicity in the quantization noise may be desirable for reducing the subjective distortion. We do not know, but that is not the problem. It is clear that the actual spectrum of noise is considerably different from the desired spectrum. Such uncontrolled differences between the actual and the desired noise spectra in a coder make it difficult to optimize the subjective performance of the coder.

[4] The differences need not be represented as the differences only in the two waveforms. One could instead compare the amplitude and phase spectra of the two speech signals and combine them in a single measure of error. Our knowledge of the exact roles of amplitude and phase spectra in speech perception is still incomplete. A quantitative model for computing the error measure in terms of differences in amplitude and phase spectra is thus not available.

Tree Encoding with a Specified Error Spectrum

Fig. 23 shows a block diagram of a tree coder for speech signals using adaptive source and error-weighting filters. The source filter is identical to the receiver of an adaptive predictive coder. It includes two feedback loops: the first loop with the *pitch* predictor $P_d(z)$ and a second loop with the *spectral envelope* predictor $P_s(z)$. The two predictors are determined using the procedure described in Section II. The excitation (*innovation sequence*) v_n for the source filter is assumed to be a sequence of independent Gaussian random numbers (zero mean and unit variance). The samples of the excitation signal are scaled by a factor σ and filtered by the source filter. The output of the source filter is compared with the original speech signal to form a difference signal which is then filtered by the error-weighting filter $1 - R$. The optimum sequence v_n is selected so as to minimize the mean-squared weighted error. The averaging interval used in computing the mean-squared weighted error is primarily determined from perceptual considerations. This interval is typically in the range of 5-15 ms.

Since v_n is a unit variance sequence, the scale factor σ is necessary to produce an optimum match between the original and the synthetic speech signals. The magnitude of σ is determined both by the power of the speech signal and the expected distortion level in the synthetic speech signal. According to the rate-distortion theory, the power of the coded signal is always smaller than the power of the signal to be coded by an amount equal to the minimum value of the mean-squared error. The optimum scale factor for white Gaussian signals is given by

$$\sigma = \max\,[0, (E_s - E_n)^{1/2}] \tag{23}$$

where E_s and E_n are the powers in the signal and noise, respectively. The intuitive meaning of (23) is that for $E_n \geqslant E_s$, no information need be sent to the receiver, because a maximum error of E_s is incurred by replacing the source with zeros. Equation (23) follows directly from two important observations: 1) that $E_n \leqslant E_s$, and 2) that the noise must be uncorrelated with the coded signal. For nonstationary signals like speech, rate-distortion theory consideration suggests that the scale factor be both frequency dependent and time varying [10]. However, a time-varying frequency-dependent scale factor can introduce undesirable characteristics in the speech signal unless the missing frequency components are filled in artificially based on the *a priori* information already available at the receiver. To avoid this problem, we have used a time-varying but frequency-independent scale factor in the initial studies. In computing the scale factor σ from (23), a knowledge of the expected distortion level in the coded signal is required. This distortion level was computed by determining the minimum possible mean-squared error (based on rate distortion theory) for a Gaussian source with a spectrum equal to the short-time spectrum of the speech signal in a given frame [30].

Tree Search Strategies

We restrict our discussion in this section to binary trees, i.e., *trees* with *two* branches from each node. A list of 1000 independent random Gaussian numbers were generated once and stored both at the transmitter and the receiver. The branches of the binary tree were "populated" with these numbers as needed in a sequential fashion. Thus, the first branch was populated with the first random number, the second branch with the second random number, and so on. After all the 1000 random numbers were exhausted, the next branch was populated with the first random number and so on.

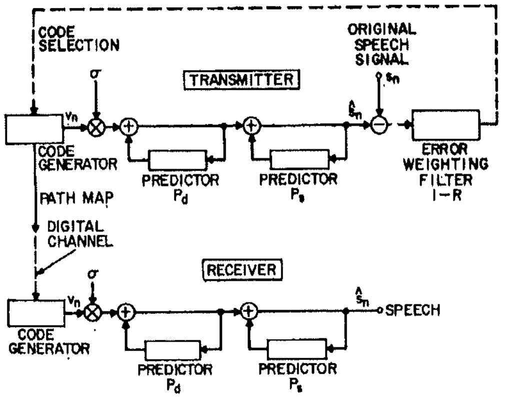


Fig. 23. Block diagram of a tree coder for speech signals using adaptive source and error-weighting filters.

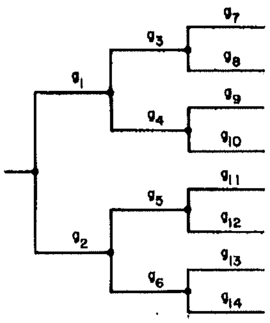


Fig. 24. Code tree populated with Gaussian random numbers.

In the above construction of the tree, each branch is populated with a single random number resulting in a rate of 1 bit/sample. Other bit rates are possible by combining different numbers of branches and random numbers per branch. An example of a binary code tree, populated with Gaussian random numbers, is shown in Fig. 24. There are only two branches at the first sample, but they increase to four at the second sample, to eight at the third sample, and so on. At each sample, one could either move to the upper branch or to the lower branch. The tree path is specified by a path map consisting of a +1 to indicate movement to the upper branch and a −1 to indicate movement to the lower branch. In the code tree shown in Fig. 24, there are a total of eight possible paths at the third sample. The resulting innovation sequences are (g_1, g_3, g_7), (g_1, g_3, g_8), (g_1, g_4, g_9), (g_1, g_4, g_{10}), (g_2, g_5, g_{11}), (g_2, g_5, g_{12}), (g_2, g_6, g_{13}), and (g_2, g_6, g_{14}). Each innovation sequence is uniquely associated with one of the eight binary path maps. The impulse responses of both the source filter H and the error-weighting filter W last over a fairly long time. Consequently, the full contribution of a particular sample in the innovation sequence does not appear in the total error until many samples later. In a tree search proce-

Exhibit T Page 412

dure, the decision to select a particular branch at the sampling instant n is made L samples later, that is, at the sampling instant $n + L$. The parameter L is thus the encoding delay. An exhaustive search to determine the optimal path map in a tree is usually impractical except for very short encoding delays, because the number of paths which must be searched increases exponentially with the encoding delay. From perceptual considerations, the desirable value of L is typically in the range 40-120 at a sampling frequency of 8 kHz. The efficiency of a variety of tree-searching algorithms has been investigated by Anderson [29]. One particularly simple yet effective procedure is the so-called M-algorithm. It progresses through the tree one level at a time and a maximum of only M lowes distortion paths are retained at each level. At the next level, the $2M$ extensions of these M paths are compared and the worst M paths are eliminated. This process is continued until the level L is reached. At that point, the accumulated error over the past L samples is examined and the best path which minimizes the error is determined. The branch L levels earlier in the best path is released and the corresponding binary symbol (indicating whether this branch is reached by an up or down motion from the previous branch in the tree) is sent to the receiver. All paths originating from the other branch (there are two branches at every level in a binary tree) are pruned. The process is repeated at the next level by accumulating the mean-squared error over the previous L samples. The error accumulation is, of course, done recursively by adding the contribution of the squared error at the new level and by subtracting the contribution from the released branch. The amount of computation grows only linearly with M in the M-algorithm. Thus, it is a computationally efficient procedure. We find that M should be at least 64 to provide reliable identification of the optimum path. The tree search procedure still requires a fairly large storage; the memory of both the source and the error-weighting filters must be saved for all the paths which are examined for minimum distortion.

An example of the waveforms of original and coded speech using a binary tree with $M = 64$ and $L = 60$ is shown in Fig. 25. The correspondence between the two waveforms is very close. The segmental SNR and speech power for a sentence-length speech utterance spoken by a male speaker is shown in Fig. 26. Informal listening tests with several sentences spoken by both male and female speakers indicate that the reconstructed speech signal has no audible noise. Only in close headphone listening can one detect that the reconstructed speech signal is slightly different (although not distorted) from the original speech signal. In another test, the innovation sequence v_n was generated by a uniform random number generator. Both the objective SNR and the perceived speech quality were almost identical for the Gaussian and uniform distributions.

We have also investigated the effect of varying the encoding delay L on the subjective performance of the coder. The speech quality is only slightly inferior with $L = 30$, but shows noticeable distortions at much lower values of L.

There are many different ways of populating the branches of a tree. We do not yet know the optimum procedure. In the tree code discussed earlier, only the *path map* was binary, but not the innovation sequence. Each path map was associated

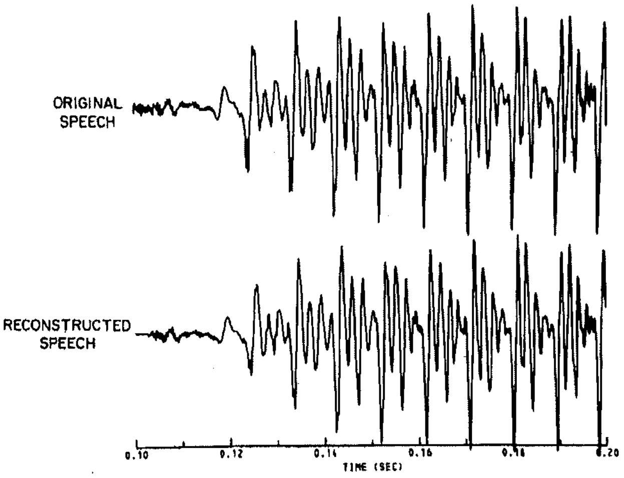


Fig. 25. Example of the waveforms of original and coded speech signals using a binary tree (1 bit/sample) with $M = 64$ and $L = 60$.

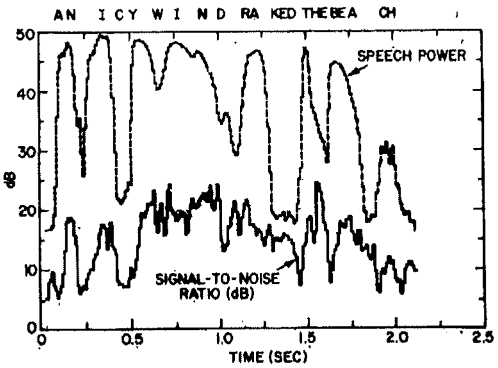


Fig. 26. Segmental SNR obtained with the tree coder shown in Fig. 23 for an utterance spoken by a male speaker.

with a unique innovation sequence. Furthermore, the branches were populated with Gaussian random numbers. We will call such a tree a stochastic tree. As an alternative, one could populate each upper branch with a +1 and each lower branch with a −1. Such a tree produces binary innovation sequences. We find the subjective performance of such a binary tree inferior to the tree populated with either Gaussian or uniform random numbers even with large encoding delays.

Our results on tree encoding are very preliminary so far. More studies are needed to determine the optimum strategies for populating the tree branches and the interactions between the subjective performance and the different parameters, such as the encoding delay and the maximum number of paths kept open in the search procedure. These early results do indicate that the tree encoding with adaptive source and error-weighting filters is potentially a very promising approach towards achieving high speech quality at low bit rates.

VI. CONCLUDING REMARKS

We have discussed in this paper further generalizations of predictive coders for speech coding at low bit rates. Waveform coders are traditionally thought to be suitable only for speech coding at medium to high bit rates. Speech coding at low bit rates has been largely left for a long time to vocoders and their derivatives. Recent work on predictive coding has demon-

Exhibit T Page 413

strated that waveform coders have the potential of providing superior performance even at low bit rates. This paper has emphasized the importance of minimizing the perceptual distortion in speech coders. The objective SNR, which has been a commonly used measure for evaluating waveform coders, becomes largely irrelevant in determining speech quality at low bit rates. Indeed, future progress in improving the speech quality in low-bit-rate coders will come primarily from recognizing what we hear and what we do not.

Delayed (tree) coding when combined with adaptive source and error-weighting filters offers an attractive framework for optimizing the performance of speech coders at any given bit rate. It is in fact an analysis-by-synthesis approach to speech coding which is very flexible and allows easy incorporation in the coder of any new understanding gained either in speech perception or generation.

REFERENCES

[1] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.

[2] J. B. O'Neal, Jr., "Predictive quantizing systems (differential pulse code modulation) for the transmission of television signals," *Bell Syst. Tech. J.*, vol. 45, pp. 689–721, May–June 1966.

[3] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. Conf. Commun., Processing*, Nov. 1967, pp. 360–361.

[4] ——, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, pp. 1973–1986, Oct. 1970.

[5] ——, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, June 1979.

[6] R. E. Crochiere and J. M. Tribolet, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, Oct. 1979.

[7] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, Dec. 1979.

[8] ——, "Objective measure of certain speech signal degradations based on properties of human auditory perception," in *Frontiers of Speech Communication Research*, B. Lindblom and S. Ohman; Eds. London, England: Academic, 1979, pp. 217–229.

[9] J. Makhoul and M. Berouti, "Adaptive noise spectral shaping and entropy coding in predictive coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 63–73, Feb. 1979.

[10] M. R. Schroeder and B. S. Atal, "Rate distortion theory and predictive coding," in *Proc. Int. Conf. Acoust., Speech, Signal processing*, Atlanta, GA, Mar. 1981, pp. 201–204.

[11] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction,," *J. Acoust. Soc. Amer.* vol. 50, pp. 637–655, Aug. 1971.

[12] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[13] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[14] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309–321, June 1975.

[15] A. H. Gray, Jr. and J. D. Markel, "Quantization and bit allocation in speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 459–473, Dec. 1976.

[16] F. Itakura, "Optimal nonlinear transformation of LPCs to improve quantization properties," *J. Acoust. Soc. Amer.*, vol. 56 (suppl.), paper H14, p. 516, 1974.

[17] P. Noll, "On predictive quantizing schemes," *Bell Syst. Tech. J.*, vol. 57, pp. 1499–1532, May–June 1978.

[18] E. G. Kimme and F. F. Kuo, "Synthesis of optimum filters for a feedback quantization system," *IEEE Trans. Circuit Theory*, vol. CT-10, pp. 405–413, Sept. 1963.

[19] B. S. Atal and M. R. Schroeder, "Optimizing predictive coders for minimum audible noise," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Washington, DC, Apr. 1979, pp. 453–455.

[20] N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM quantizers," *Proc. IEEE*, vol. 62, pp. 611–632, May 1974.

[21] B. S. Atal and M. R. Schroeder, "Improved quantizer for adaptive predictive coding of speech signals at low bit rates," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Denver CO, Apr. 1980, pp. 535–538.

[22] F. Jelinek and K. S. Schneider, "On variable-length-to-block coding," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 765–774, Nov. 1972.

[23] D. Pan, "Quantization and channel encoding of the APC speech prediction residual," thesis, Dept. Elec. Eng., Massachusetts Inst. Technol., Cambridge, May 1981.

[24] F. Jelinek, "Tree encoding of memoryless time-discrete sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 584–590, Sept. 1969.

[25] N. S. Jayant and G. A. Christensen, "Tree encoding of speech using the (M,L)-algorithm and adaptive quantization," *IEEE Trans. Commun.*, vol. COM-26, pp. 1376–1379, Sept. 1978.

[26] J. B. Anderson and J. B. Bodie, "Tree encoding of speech," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 4, pp. 379–387, 1975.

[27] H. G. Fehn and P. Noll, "Tree and trellis coding of speech and stationary speech-like signals," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, Apr. 1980, pp. 547–551.

[28] S. G. Wilson and S. Husain, "Adaptive tree encoding of speech at 8000 bits/s with a frequency-weighted error criterion," *IEEE Trans. Commun.*, vol. COM-27, pp. 165–170, Jan. 1979.

[29] F. Jelinek and J. B. Anderson, "Instrumentable tree encoding of information sources," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 118–119, Jan. 1971.

[30] R. A. McDonald and P. M. Schultheiss, "Information rates of Gaussian signals under criteria constraining the error spectrum," *Proc. IEEE*, vol. 52, pp. 415–416, Apr. 1964.

★

Bishnu S. Atal (M'76–SM'78–F'82) was born in Kanpur, India, on May 10, 1933. He received the B.Sc. (honors) degree in physics from the University of Lucknow, Lucknow, India, in 1952, the Diploma in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1955, and the Ph.D. degree in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1968.

From 1957 to 1960 he was a Lecturer in Acoustics at the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore. In 1961 he came to the United States to join the Research Staff of Bell Laboratories, Murray Hill, NJ. At Bell Laboratories his work has covered a wide range of topics in acoustics such as computer simulation of sound transmission in rooms, new measurement techniques for concert halls, fading in mobile radio, automatic speaker recognition, and speech coding. More recently his research interests have centered on new methods for analysis and synthesis of speech signals. He is the author of a number of technical papers in architectural acoustics and speech communication, and holds several patents for inventions in these fields.

Dr. Atal is a Fellow of the Acoustical Society of America. He received the 1975 IEEE Acoustics, Speech, and Signal Processing Society Technical Achievement Award for fundamental contributions to linear predictive coding of speech signals. In 1980 he received, jointly with M. R. Schroeder, the IEEE ASSP Senior Award for their paper on predictive coding of speech signals and subjective error criteria.

Exhibit I Page 414